计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(10):190−195 [doi: 10.15888/j.cnki.csa.007097]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

# 基于神经网络与注意力机制的中文文本校对方法①

郝亚男, 乔钢柱, 谭　瑛

(太原科技大学 计算机科学与技术学院, 太原 030024)
通讯作者: 乔钢柱, E-mail: qiaogangzhu@sohu.com

**摘　要**: 中文文本校对是中文自然语言处理方面的关键任务之一, 人工校对方式难以满足日常工作的数据量需求, 而基于统计的文本校对方法不能灵活的处理语义方面的错误. 针对上述问题, 提出了一种基于神经网络与注意力机制的中文文本校对方法. 利用双向门控循环神经网络层获取文本信息并进行特征提取, 并引入注意力机制层增强词间语义逻辑关系的捕获能力. 在基于 Keras 深度学习框架下对模型进行实现, 实验结果表明, 该方法能够对含语义错误的文本进行校对.

**关键词**: 中文文本校对; 注意力机制; 双向门控循环神经网络; 端到端序列模型

引用格式: 郝亚男,乔钢柱,谭瑛.基于神经网络与注意力机制的中文文本校对方法.计算机系统应用,2019,28(10):190–195. http://www.c-s-a.org.cn/1003-3254/7097.html

## Chinese Text Proofreading Method Based on Neural Network and Attention Mechanism

HAO Ya-Nan, QIAO Gang-Zhu, TAN Ying

(School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China)

**Abstract**: Chinese text proofreading is one of the key tasks in Chinese natural language processing, and manual proofreading is difficult to meet the data volume requirement of daily work, and the text proofreading method based on statistics can not deal with semantic errors flexibly. Aiming at the above problems, a Chinese text proofreading method based on neural network and attention mechanism is proposed. The bidirectional Gated Recurrent Unity neural network layer is used to obtain text information and feature extraction, and the ability of attention mechanism layer to enhance the semantic logic relation between words is introduced. The model is implemented under the framework of deep learning based on Keras. Experimental results show that this method can proofread text with semantic errors.

**Key words**: proofreading of Chinese text; attention mechanism; bidirectional Gated Recurrent Unity (GRU); end-to-end sequence model

自然语言 OCR 识别后文本错误自动校对, 已经引起越来越多的关注. 近年来, 我国法治建设的快速发展, 类型多样的法律案件数量增多. 由于实际情况所限制, 我国司法机关处理的案件卷宗以纸质卷宗为主, 想要在较短的时间内获取有效的信息, 较为困难. 随着信息技术的广泛普及, 我国已逐渐将电子卷宗应用辅助办案系统中. 为加快纸质卷宗电子化, 电子化过程中采取 OCR 识别技术. 但是由于纸质卷宗的打印质量低或扫描不当等原因, 导致纸质卷宗 OCR 识别效果不好. 因此, 在电子卷宗应用于后续任务前, 需要有效的校对器

① 基金项目: 山西省重点研发计划重点项目 (201703D111011)
Foundation item: Major Program of Key Research and Development Program of Shanxi Province (201703D111011)
收稿时间: 2019-03-20; 修改时间: 2019-04-17; 采用时间: 2019-04-19; csa 在线出版时间: 2019-10-15

来帮助纸质卷宗 OCR 识别后的文本自动校对.

由于中文文本与英文文本特点不同, 中文文本校对是在错误文本的字词、语法或语义等来进行校对的. 目前, 针对字词级的 OCR 识别后的中文文本校对研究相对比较充分, 但在 OCR 识别后的中文文本还存在许多其他类型错误, 这些错误从字词级的角度来看, 可能不存在问题, 但是不符合当前文本中的上下文语义搭配, 例如: “透过中间人向另一方表示无欠债关系.”. 其中, “透过”就是不符合文本语义搭配, 此处应表示为“通过”. 因此, 本文主要是研究如何结合语义校对中文文本中的错误.

## 1 相关工作

在 20 世纪 60 年代起, 国外就开始对英文文本拼写自动校对进行研究. 在研究初期主要是建立语言模型与字典来进行字词级[1–4]的校对. 近年来, 字词级校对的研究已经较充分, 但在真词错误校对时, 若不限制给定语境, 那文本校对的可靠性就难以保证. 因此, 学者们在基于语义对文本校对展开进一步研究.

Hirst[5]等在文本校对的计算中加入语义信息, 采用 WordNet 来计算词与词间的语义距离, 若词间语义距离较远, 则判断这个词是错误的, 反之, 若发现与上下文距离语义较近的词就可能被作为正确的词. Kissos 等[6]是基于 OCR 识别后的阿拉伯语校对, 其采取的方式是通过与混淆矩阵相结合的语料库形成候选数据集; 然后对每个单词所提取的特征对单词分类, 将候选集中排名最高的单词作为校对建议. Sikl[7]等将校对问题看作翻译问题来解决, 把错误文本作为被原语言, 纠正文本作为目标语言进行文本拼写纠错. 张仰森等[8]提出了一种基于语义搭配知识库和证据理论的语义错误侦测模型, 构建三层语义搭配知识库以及介绍了基于该知识库和证据理论的语义侦测算法, 有效地进行语义级错误侦测. Konstantin 等人[9]提出基于边际分布和贝叶斯网络计算的方法, 在一定程度上提高了低质量图像的文档字段 OCR 识别后校对准确率. 陶永才等[10]基于构建词语搭配知识库, 综合使用互信息和聚合度评价词语关联强度, 进行词语搭配关系校对. Liu 等人[11]提出基于注意机制的神经语法检错模型, 将解码端视为二进制分类器进行语法检错. 刘亮亮等人[12]面向非多字词错误提出基于模糊分词的自动校对方法. 姜赢等人[13]提出基于描述逻辑本体推理的语义级中文校对方法, 通过描述逻辑推理机来判断提取的语义内容的逻辑一致性, 并将检测出的逻辑一致性错误映射为中文语义错误. Xie 等人[14]通过具有注意机制的编码器和解码器的递归神经网络来进行字级别的英文文本校对. Yu 等人[15]通过语言模型、拼音及字形完成校对工作.

分析以上文献发现, 在以往中文文本自动校对的方法中均进行了大量知识的准备工作, 知识库的完善程度对校对结果有很大影响. 为了减少知识库等相关知识对校对效果的影响, 采用深度学习模型的思路完成文本自动校对任务. 通过模型的自学习获取词间相关性, 来完成文本校对任务, 在一定程度上减少了人为干预. 模型采用端到端序列模型, 在解码端与编码端构成成分的选择上, 主要是从时间方面考虑选取了门控循环网络与注意力机制层结合构成, 最后通过 Dense 层和 Softmax 层完成文本自动校对任务.

## 2 基于神经网络与注意力机制的校对模型

### 2.1 门控循环神经网络

长短期记忆网络 (Long Short-Term Memory, LSTM) 在自然语言处理任务中有着广发的应用, 但 LSTM 在训练耗时长、参数多等问题. 研究人员在 2014 年对 LSTM 进行优化调整, 提出了门控循环神经网络 (Gated Recurrent Unity, GRU).

GRU 保持 LSTM 优点的同时又使得内部结构更加简单. GRU 由更新门和重置门两个门组成, 更新门用于控制前一时刻的隐层输出对当前时刻的影响程度, 更新门的值越大说明前一时刻的隐藏状态对当前时刻隐层的影响越大; 重置门用于控制前一时刻隐层状态被忽略的程度, 重置门值越小说明被忽略的越多.

GRU 的结构如图 1 所示.

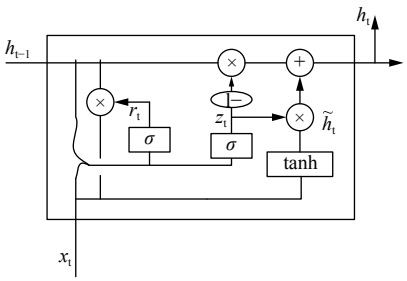


图 1 GRU 神经元结构图

GRU 的更新方式如式 (1) 至式 (5) 所示:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad (1)$$

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{2}$$

$$\tilde{h}_t = \tanh(W_{\tilde{h}} \cdot [r_t * h_{t-1}, x_t]) \tag{3}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{4}$$

$$y_t = \sigma(W_o \cdot h_t) \tag{5}$$

其中, $r_t$ 和 $z_t$ 分布表示为 $t$ 时刻的重置门和更新门, $\tilde{h}_t$、$h_t$分别表示 $t$ 时刻的候选激活状态、激活状态. $h_{t-1}$ 为 $(t-1)$ 时刻的隐层状态.

### 2.2 双向门控循环神经网络 (BiGRU)

BiGRU 能够同时将当前时刻的输出同前一个时刻的状态与后一时刻的状态产生联系. BiGRU 是由单向、方向相反、当前时刻的输出由方向相反的两个 GRU 输出共同决定的神经网络模型. BiGRU 的结构模型如图 2 所示.

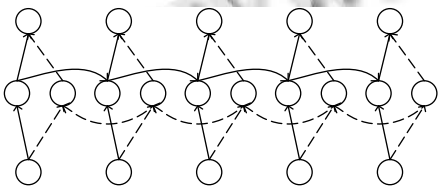

图 2 BiGRU 结构模型图

在 $t$ 时刻 BiGRU 的隐层状态计算公式如式 (6) 至式 (8) 所示:

$$\overrightarrow{h}_t = GRU(x_t, \overrightarrow{h_{t-1}}) \tag{6}$$

$$\overleftarrow{h} = GRU\left(x, \overleftarrow{h}\right) \tag{7}$$

$$h_t = w_t \overrightarrow{h_t} + v_t \overleftarrow{h_t} + b_t \tag{8}$$

其中, GRU() 表示对输入词向量的非线性变换, 将词向量编码为 GRU 隐层状态. $w_t$ 为 $t$ 时刻前向隐层状态 $\overrightarrow{h_t}$对应的权重, $v_t$ 为 $t$ 时刻反向隐层状态$\overleftarrow{h_t}$对应的权重, $b_t$ 为 $t$ 时刻隐层状态对应的偏置.

### 2.3 注意力机制

注意力机制就是通过对关键部分加强关注、突然局部重要信息, 简单来说就是计算不同时刻数据的概率权重, 突出重点词语. 多头注意力机制[16]将序列分为 key, values 和 query. 多头注意力机制通过尺度化的点积方式并行多次计算, 每个注意力输出是简单拼接、线性转换到指定的维度空间而生成的. 多头注意力机制结构如图 3 所示.

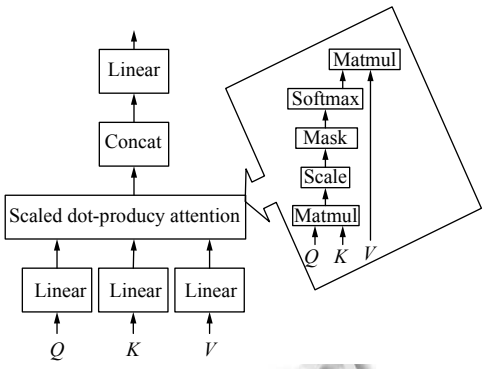


图 3 多头注意力机制结构图

多头注意力机制层可以视为一个序列编码层, 从初始隐层状态到新隐层状态 $z$ 的计算公式如式 (9) 所示.

$$z_i = \sum_{j=1}^{n} a_{ij}(x_i W^V) \tag{9}$$

其中, 权重系数 $a_{ij}$ 的计算公式如式 (10) 所示.

$$a_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{n} \exp e_{ik}} \tag{10}$$

其中, $e_{ij}$ 的计算公式如式 (11) 所示.

$$e_{ij} = \frac{\left(x_i W^Q\right)\left(x_j W^K\right)^{\mathrm{T}}}{\sqrt{d_z}} \tag{11}$$

选择了可扩展的点积来实现兼容性功能, 从而实现高效的计算. 输入的线性变换增加了足够的表达能力. $W^Q, W^K, W^V$ 是参数矩阵, $W^Q, W^K, W^V \in \mathbb{R}^{dx \times dz}$.

### 2.4 基于神经网络与注意力机制的校对模型

对于文本采用生成的方式进行校对, 首先句子是由字、词和标点组成的有序的序列, 若对句中某个字词进行纠正, 则需要通过上下文信息进行推断和生成. 在中文文本校对的研究中, 仅使用神经网络抓取的上下文特征信息作为语义校对是不够的. 上下文信息对当前字词的校对影响力不同, 不能作为同一标准对当前字词的校对产生作用. 因此, 本文构建了一个基于注意力机制的序列到序列的中文文本校对模型. 模型引入基于注意力的神经网络, 以增强获取词与词间的依赖性的能力.

整体模型架构如图 4 所示.

#### 2.4.1 文本向量化

模型进行文本校对时, 首先要将文本向量化. 通过一个特定维度的向量代表词, 词向量可以刻画词与词在语义上的相关性, 并将词向量作为神经网络的输入.

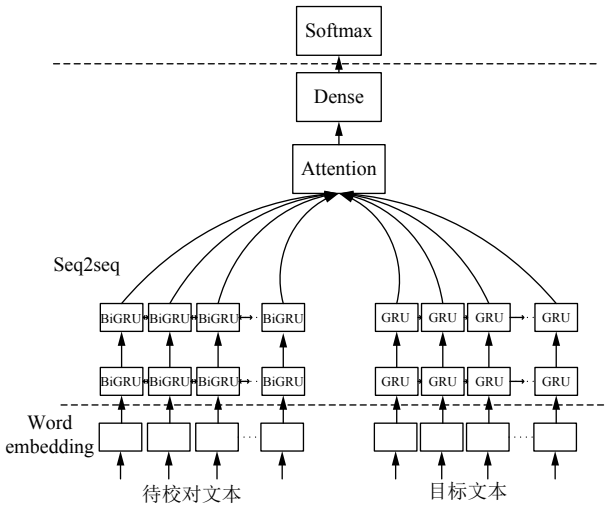


图 4 模型架构

将训练语料、测试语料集以及开始标志等所有字词建立一个大小为 $N$ 的词字典矩阵, $N$ 表示字典的大小. 建立一个词到词字典的映射关系查找表, 将输入的词转换为序号, 之后将序号转换为词嵌入向量.

2.4.2 序列到序列端

模型的编码端由 BiGRU 层构成, 文本向量化后的词向量作为 BiGRU 层的输入. BiGRU 层主要目的是对输入的待校对文本进行特征提取. 正向 GRU 通过从左向右的方式读取输入的待校对句子 $X$, 从而得到正向的隐层状态序列$\left\{\overrightarrow{h}_1,\overrightarrow{h}_2,\cdots,\overrightarrow{h}_i\right\}$. 反向 GRU 是从右往左的方式读取输入的待校对句子 $X$, 同样可以得到反向隐层状态序列$\left\{\overleftarrow{h}_1,\overleftarrow{h}_2,\cdots,\overleftarrow{h}_i\right\}$. 将正向和方向隐层状态序列进行连接得到编码端的隐层状态序列 $h$, 其中:

$$h_i=[\overrightarrow{h}_i;\overleftarrow{h}_i] \tag{12}$$

模型的解码端是采用单向 GRU, 每一时刻的隐层状态 $w_i$ 均由前一时刻的隐层状态 $w_{i-1}$ 和上一时刻的输出 $y_{i-1}$.

$$w_i=GRU(w_{i-1},y_i) \tag{13}$$

注意力机制层通过计算输入序列 $x_1$, $x_2$, $x_3$, …, $x_n$ 每个字词对于 $i$ 时刻输出值 $y_i$ 的影响权重加权求和所得. 在生成校对结果时, 解码信息融合了输入序列对输出序列每个时刻的概率分布.

2.4.3 基于集束搜索的校对算法

采用集束搜索 (beam search) 求解校对位置的最优结果. 基于集束搜索的校对算法如算法 1 所示.

算法 1. 基于集束搜索的校对算法

```
#x_i 为待校对句子
#proba 用来记录候选词 y_i 以及得分 score, #beam 的值设置为 N
# max_target_len 为目标句子的最大长度
for i in rang(max_target_len) #predict 根据 x_i 预测所有可能的字词及其得分
    proba=predict(x_i)
    #生成对所有候选集排序
    for j in len(x):
      for y_i, score in proba:
        if score>new_score:
          new_top=y_i
          new_score=score
      else:
        t=[y_i, score]
        new_beam.append(t)
#取 new_beam 中最好的 beam 个候选集 c
c=get_max(new_beam)
```

## 3 实验结果与分析

校对方法由两个阶段组成: 训练和校对. 如图 5 所示.

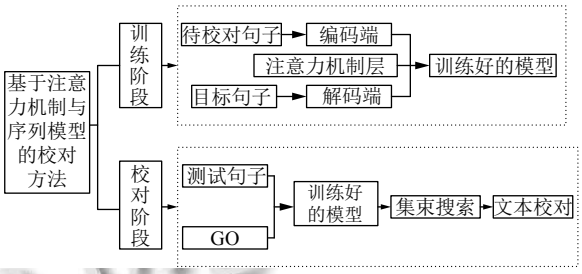


图 5 基本框架

### 3.1 实验数据

实验采用在网站中抓取的公开法律文书文本整理后的句子作为训练集, 样本数据总量为 10.7 MB, 随机抽取 404 句作为测试集. 例如, “透过中间人向另一方表示无欠债关系.”应为“通过中间人向另一方表示无欠债关系.”.

### 3.2 建模过程及参数

使用基于 Keras 的深度学习框架进行模型实现. 基于双向门控循环神经网络和注意力模型的方法已在第 2 节中介绍. 首先, 将输入句子向量化, 作为模型的输入; 其次, 添加 BiGRU 和 GRU 层, 并在 GRU 层后添加注意力机制层; 然后, 添加双层 Dense 层, 在 Dense 层采用 ReLU 激活函数. 同 Sigmoid 激活函数相比,

ReLU 激活函数能实现单侧抑制[17], 能够有效防止过拟合. 因此, 在实验中选取 ReLU 激活函数; 最后, 构建 Softmax 层对文本进行校对, 作用是将输出转变成概率, 通过输出的概率向量结合词典反向映射获得当前时刻的输出词.

在解码时, “GO”表示一个句子的开始标志, “END”表示一个句子的结束标志, “PAD”为补充长度的符号. “GO”和“END”在解码器端作为开始解码和结束解码的标志, 并一次生成一个字词直到遇到结束标志符号.

训练模型使用 Adam 优化[18], 词向量维度为 128, 每层神经元个数设置为 128, *loss* 函数采用 categorical_crossentropy.

### 3.3 实验评价标准

本文采用准确率 ($P$), 召回率 ($R$) 以及 $F_{0.5}$ 值作为实验的评价标准. 准确率反应校对结果的准确程度, 召回率表示校对结果的全面性, $F_{0.5}$ 值为准确率和召回率的综合评价的指标.

$$P = \frac{\text{正确校对个数}}{\text{校对总数}} \tag{14}$$

$$R = \frac{\text{正确校对个数}}{\text{错误总数}} \tag{15}$$

$$F_{0.5} = \frac{1.25 \times P \times R}{0.25 \times P + R} \tag{16}$$

### 3.4 实验结果分析

采用未加注意力机制的序列到序列模型做为基线模型 (baseline), 实验将本文提出的模型 (BiGRU-A-GRU) 与基线模型以及其他模型进行对比, 在同一数据集上进行训练和测试, 得到的中文文本校对的实验结果如表 1 所示.

表 1 基于不同校对方法的结果

| 方法 | $P$ | $R$ | $F_{0.5}$ |
|---|---|---|---|
| Baseline | 0.3100 | 0.3039 | 0.3088 |
| BiGRU-A | 0.0419 | 0.0343 | 0.0401 |
| BiLSTM-A-BiLSTM | 0.3350 | 0.3284 | 0.3337 |
| BiGRU-A-GRU | 0.3417 | 0.3334 | 0.3400 |

从表 1 中可以得到, 本文提出的模型在语义方面的中文文本校对的完成情况好于基线方法, 其准确率、召回率、$F_{0.5}$ 值均有一定的提高. 这些文本校对效果的提升主要由于 BiGRU-A-GRU 模型增强了对词间语义关系的捕捉能力, 同时该模型减少了因错误侦测产生的影响.

测试集对应最高准确率时的迭代时间为模型的迭代时间, 如表 2 所示.

表 2 模型迭代时间对比

| 方法 | Time/epoch(s) |
|---|---|
| Baseline | 1186 |
| BiGRU-Attention | 1192 |
| BiLSTM-A-BiLSTM | 2206 |
| BiGRU-A-GRU | 1660 |

BiGRU-A-GRU 模型与 BiLSTM-A-LSTM 模型相比较, 均采用了 Attention 层, 区别是一个采用了 BiGRU 层一个采用了 BiLSTM 层, 表 2 可以看出在模型迭代时, BiLSTM-A-LSTM 模型迭代用时更长.

总之, 通过表 1, 表 2 及图 6 可以得知: 在本数据集上, BiGRU-A-GRU 模型优于 BiLSTM-A-LSTM 模型, 因为 BiGRU 相比 BiLSTM 收敛速度快, 参数更少, 在一定程度上降低了模型的训练时间, Attention 层在校对过程中能对句子中关键部分加强关注, 突出相关联的词语完成校对任务.

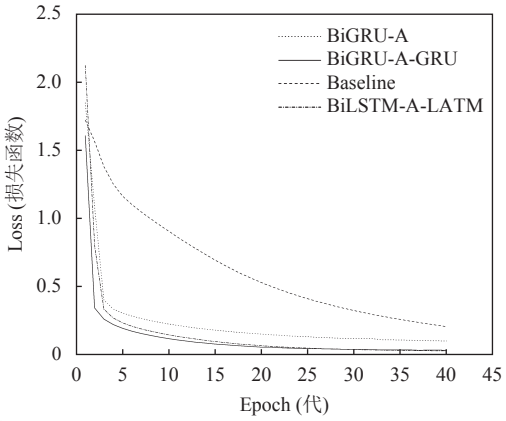


图 6 模型训练损失率变化曲线

## 4 结束语

本文提出一种基于神经网络与注意力机制的中文文本校对方法. 将注意力机制引入文本校对任务中, 捕捉词间语义逻辑关系, 提升了文本校对的准确性. 实验证明, 深度学习模型中引入注意力机制能够提高中文文本自动校对的准确性.

中文文本词语含义的多样性, 对语义错误的文本校对的发展有一定的阻碍性. 在未来工作中, 将寻找能够提高模型学习词间语义关系的途径, 进而更好地完成文本自动校对任务, 并且采用对系统的计算和开销等影响较小的方法.

## 参考文献


1 Pollock JJ, Zamora A. Automatic spelling correction in scientific and scholarly text. Communications of the ACM, 1984, 27(4): 358–368. [doi: 10.1145/358027.358048]

2 Kukich K. Techniques for automatically correcting words in text. Proceedings of 1993 ACM Conference on Computer Science. Indianapolis, IN, USA. 1993. 515.

3 张仰森, 俞士汶. 文本自动校对技术研究综述. 计算机应用研究, 2006, 23(6): 8–12. [doi: 10.3969/j.issn.1001-3695.2006.06.002]

4 马金山, 张宇, 刘挺, 等. 利用三元模型及依存分析查找中文文本错误. 情报学报, 2004, 23(6): 723–728. [doi: 10.3969/j.issn.1000-0135.2004.06.014]

5 Hirst G, Budanitsky A. Correcting Real-word Spelling Errors by Restoring Lexical Cohesion. New York: Cambridge University Press, 2005.

6 Kissos I, Dershowitz N. OCR error correction using character correction and feature-based word classification. Proceedings of the 2016 12th IAPR Workshop on Document Analysis Systems. Santorini, Greece. 2016. 198–203.

7 Siklósi B, Novák A, Prószéky G. Context-aware correction of spelling errors in Hungarian medical documents. Proceedings of the 1st International Conference on Statistical Language and Speech Processing. Tarragona, Spain. 2013. 248–259.

8 张仰森, 郑佳. 中文文本语义错误侦测方法研究. 计算机学报, 2017, 40(4): 911–924.

9 Bulatov K, Manzhikov T, Slavin O, *et al*. Trigram-based algorithms for OCR result correction. Proceedings of SPIE 10341, Ninth International Conference on Machine Vision. Nice, France. 2016. 1034100. [doi: 10.1117/12.2268559]

10 陶永才, 海朝阳, 石磊, 等. 中文词语搭配特征提取及文本校对研究. 小型微型计算机系统, 2018, 39(11): 2485–2490. [doi: 10.3969/j.issn.1000-1220.2018.11.025]

11 Liu ZR, Liu Y. Exploiting unlabeled data for neural grammatical error detection. Journal of Computer Science and Technology, 2017, 32(4): 758–767. [doi: 10.1007/s11390-017-1757-4]

12 刘亮亮, 曹存根. 中文“非多字词错误”自动校对方法研究. 计算机科学, 2016, 43(10): 200–205. [doi: 10.11896/j.issn.1002-137X.2016.10.038]

13 姜赢, 庄润钹, 吴烨凡, 等. 基于描述逻辑本体推理的语义级中文校对方法. 计算机系统应用, 2017, 26(4): 224–229. [doi: 10.15888/j.cnki.csa.005680]

14 Xie ZA, Avati A, Arivazhagan N, *et al*. Neural language correction with character-based attention. arXiv: 1603.09727, 2016.

15 Yu JJ, Li ZH. Chinese spelling error detection and correction based on language model, pronunciation, and shape. Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing. 2014. 220–223.

16 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach, CA, USA. 2017.

17 Yarotsky D. Error bounds for approximations with deep ReLU networks. Neural Networks, 2017, 94: 103–114. [doi: 10.1016/j.neunet.2017.07.002]

18 Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv: 1412.6980, 2014.